Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 177–180
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

STS-korpus: A Sign Language Web Corpus Tool for Teaching and Public Use

Zrajm Öqvist, Nikolaus Riemer Kankkonen, Johanna Mesch
Department of Linguistics, Stockholm University
SE-106 91 Stockholm, Sweden
{zrajm, nikolaus.kankkonen, johanna.mesch}@ling.su.se

Abstract

In this paper we describe *STS-korpus*, a web corpus tool for Swedish Sign Language (STS) which we have built during the past year, and which is now publicly available on the internet. *STS-korpus* uses the data of Swedish Sign Language Corpus (SSLC) and is primarily intended for teachers and students of sign language. As such it is created to be simple and user-friendly with no download or setup required. The user interface allows for searching – with search results displayed as a simple concordance – and viewing of videos with annotations. Each annotation also provides additional data and links to the corresponding entry in the online Swedish Sign Language Dictionary. We describe the corpus, its appearance and search syntax, as well as more advanced features like access control and dynamic content. Finally we say a word or two about the role we hope it will play in the classroom, and something about the development process and the software used. *STS-korpus* is available here: **https://teckensprakskorpus.su.se**

Keywords: web application, user-focused, accessibility, language teaching, Swedish Sign Language corpora

1. Introduction

Corpora are a valuable tool in second language teaching (Granath, 2009, 245) – however downloading and installing the software and data needed for using a corpus can be a daunting task for students, teachers, and researchers alike.

This is especially relevant in sign language courses for beginners, where no prior knowledge of corpora is expected. These classes would benefit from corpus use, but walking students through a time consuming and complicated installation process takes time away from the primary purpose of the course – language teaching. Thus, there is need for a simpler tool, one which requires less instruction, and has less initial setup, and this is where *STS-korpus* enters the picture (*STS-korpus*, 2020).

For the past year we have been developing *STS-korpus*, a web corpus interface, meaning that we can now simply tell our students to go to **https://teckensprakskorpus.su.se** where they can immediately access our corpora, without the need for them to install anything.

2. Role in Teaching

It is our hope that *STS-korpus* will be used among teachers and students of sign language, both inside and outside of classroom.

STS-korpus was designed specifically with the learning situation in mind, as a simple and fast lookup tool, for use in situations where downloading, installing and configuring a full corpus is beyond the scope of the current exercise, regardless of whether this is because of a lack in technical expertise, available hard drive space, or lack of admin privileges on a classroom computer.

We hope to facilitate the teacher's role by providing easy access to a corpus. This can be used both for making presentations during class, and also as an aid in answering student questions.

The role of the corpus in sign language teaching is important since sign language does not have a written form. Therefore a corpus is the only way to look up how a particular sign is used in everyday conversation.

At the end of the development process we informally reached out to a few teachers of sign language for their opinions. They appreciated the possibilities of the *STS-korpus*, especially the easy availability. This leads us to think that we are on the right track, and that *STS-korpus* can have a future in education.

2.1. In Relation to ELAN

SSLC (Mesch et al., 2012) and later corpora created here at the Department of Linguistics at Stockholm University were all annotated using ELAN (2020). We will therefore here contrast *STS-korpus* with ELAN, rather than iLex (Hanke, 2002) or other tools.

STS-korpus and ELAN have different purposes, and are complementary to each other.

It is only a slight exaggeration to say that ELAN is the Swiss army knife of annotation, a tool that can do almost anything. It can be used for annotation, searching, making statistics and more (Crasborn and Sloetjes, 2010) – everything except real-time collaboration. However it is a complex tool which takes time to learn (Leeson et al., 2019, 345) and for someone who is new to ELAN setting up a corpus for searching and viewing is not a trivial matter.

A course on sign language corpus linguistics would probably not be complete without at least an introduction to ELAN, but for a beginner's course in sign language, introducing a program of ELAN's complexity (Mesch and Wallin, 2008) could be counterproductive.

STS-korpus, on the other hand, is more like an IKEA Allen wrench – designed to do one thing, and to do that with as little fuss as possible. This is far better suited for a beginner's course.

3. Usage

STS-korpus works on both computers and portable devices, though screen size will affect the experience, and a tablet will serve you better than a cell phone.

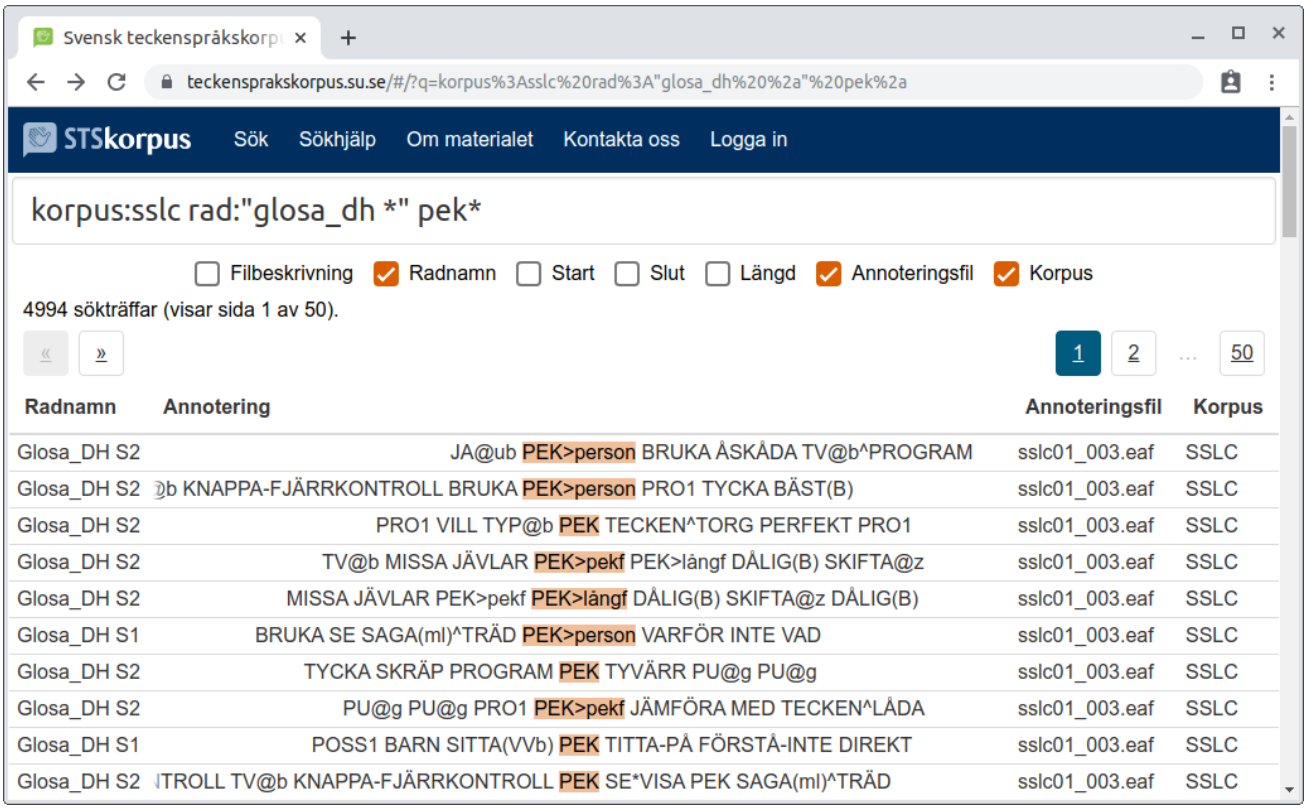

Figure 1: A search with its results

3.1. Searching

STS-korpus has a simple, uncluttered search interface (Figure 1) which is intentionally similar to that of Google and other search engines, so as to feel familiar for both novice and expert computer users.

The user may search the annotations in the database for words, or parts of words (using `*` to match the varying part). Search prefixes may also be used to limit a search to named annotation tiers (`rad:`), files (`fil:`) or corpora (`korpus:`).

Figure 1 shows the result of a search of the Swedish Sign Language Corpus (`korpus:sslc`) in tiers for the dominant hand (`rad:"glosa_dh *"`) for annotations containing pointing (`pek*`).

The tiers available depend on the corpus. SSLC tiers include: `Glosa_DH S1` and `Glosa_DH S2` (glosses for dominant hand of subject 1 and 2), `Glosa_NonDH S1` and `Glosa_NonDH S2` (glosses for the non-dominant hand), and `Översättning S1` and `Översättning S2` (Swedish translations) (cf. Figure 2).

Search results are displayed in a simple concordance view, with each found search term highlighted. The checkboxes can be used to show additional columns of information, such as file names, file descriptions, tier names, corpus names, and start/end times and length of the annotations. (The search results are currently not sorted, but we hope to add options for this in the near future.)

You can also share your result with other people by just copying the web address and sharing that with them. (However, if you are a logged-in registered user, keep in mind that you might have access to information that your recipient will not be able to see [section 4.1.].)

A description of the glossing conventions used in SSLC (Figure 1) can be found in English in Mesch and Wallin (2015), and in Swedish in Wallin and Mesch (to be published 2020).

3.2. Video Viewing

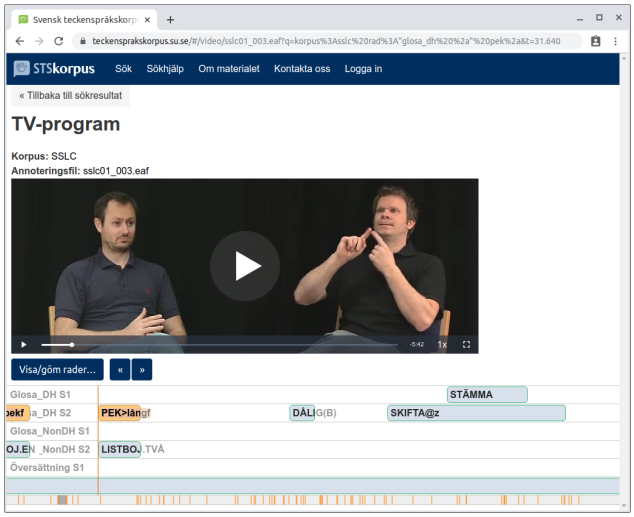

Figure 2: Video view with highlighted search "`PEK*`"

Clicking on any of the matches in the search results will load the video (Figure 2) and skip to the location of the

matching annotation. Matching annotations in the video are highlighted, and the horizontal scrollbar at the bottom of the window has marks to indicate where matches are found in the video, turning the scrollbar into a rough dispersion plot. There are buttons for selecting which annotation tiers to display, and for skipping to the previous and next match. Playback speed can also be adjusted.

This view is similar to ELAN, with annotations scrolling across the screen as the video is playing, and a thin vertical line to indicate the current position. Annotations and video are always kept in sync so that the user may navigate by either scrolling the annotations, or by skipping to a different point using the video player.

3.3. Links to Dictionary

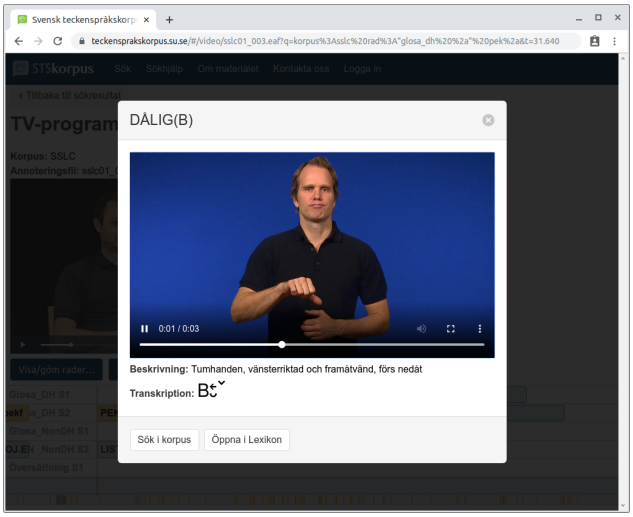

Figure 3: Video view, annotation details

There are links from *STS-korpus* to our already existing online Swedish Sign Language Dictionary (*Svenskt teckenspråkslexikon*, 2008). Clicking on a gloss annotation in the video view quickly brings up additional details about the sign, without having to loading the full dictionary entry (Figure 3). This view includes buttons for performing a new search in the corpus for this particular gloss ("Öppna i korpus") and for going to the full dictionary entry ("Öppna i lexikon").

We have also updated the Swedish Sign Language Dictionary so that each individual entry now contains a link to search for that sign in *STS-korpus*. This facilitates smooth navigation between dictionary and corpus, and makes it very easy to go from a dictionary entry to a list of real-world usage examples for a sign.

3.4. Multiple Datasets

We decided early on to build a backend which would allow for datasets from multiple corpora – a design decision which has served us well. Because of this it has been trivial to add data from other corpora, previously developed at the Department of Linguistics, as interest for the web corpus has grown.

So far we have added data from the Swedish Sign Language Corpus (SSLC) (Mesch et al., 2012), Swedish parts of the ECHO corpus project (Bergman and Mesch, 2004), as well as a special "dynamic corpus" (see section 4.3.). In the future we also hope to add parts of the Tactile Swedish Sign Language Corpus (TSSL) (Mesch, 2016) and the Corpus of Swedish Sign Language as Second Language (SSLC-L2) (Schönström and Mesch, 2017).

4. Advanced Features

4.1. Access Control

Any user can, without having to log in, access the public and anonymized part of SSLC (Mesch et al., 2012). There is also a login system in place, by which we can grant registered users additional access, e.g. access to sensitive or experimental data, as well as data from additional corpora (section 3.4.).

There are three different access levels for registered users: *teacher, researcher* and *admin*.

A *teacher* has permissions suitable for demonstrations in the classroom, i.e. access to a larger number of tiers, but not to sensitive information. A *researcher* has access to everything. And finally an *admin* has access to all data, but can also create and remove users.

4.2. Importing Data

STS-korpus needs two pieces of data for each annotated video: the annotations and the video itself.

All video files in *STS-korpus* were preprocessed, specifically scaled down and reencoded to a format suitable for the web. With the SSLC videos we also took the two camera angles (seen in Figure 2), and merged them into a single video. This was done to avoid possible sync issues while the user is viewing the video which might result from poor browser or computer performance.

Annotations are imported from the ELAN .eaf files used by SSLC, and thereafter put into separate entries in the database.

4.3. Dynamic Content

During early testing teachers at our department requested the ability to upload their own material for use in the classroom. We enabled this by setting up networked hard drive to where teachers may upload their own videos and ELAN annotation files. Files put there are automatically imported into a special corpus, `KURS` (meaning *course*), which makes them available to registered users of *STS-korpus*.

This way registered users, which have also been provided with a separate login for the drive, may upload their files. All uploaded annotations become searchable and viewable in *STS-korpus*, though, for natural reasons, only annotations with glosses that can be found in the Swedish Sign Language Dictionary will link to a dictionary entry.

This way a teacher can upload the data desired for a presentation and then, later on, find it in the web interface by adding `fil:` or `korpus:kurs` to the search.

5. Development

Programming and design of *STS-korpus* was done by Patrick Hansson and Zrajm Öqvist, working one day a week for two semesters.

Throughout the development process we have received continuous feedback from both our colleagues at the office of the Swedish Sign Language Dictionary, and, in the later stages of development, from several other researchers and sign language teachers at the Department of Linguistics.

5.1. Software Used

The frontend was written in *Javascript*, using *Vue.js* and *Buefy*.

The backend was written using *Python*, with the web server and API built using the *Flask* framework. Data is stored in a *MariaDB* database and is served to the frontend by means of a custom made JSON-based API. The database is populated from ELAN .eaf files using our own database import script, written in *Python* and making use of the *SQLAlchemy* and *pympi-ling* modules.

We decided against publishing the source code online, since it is specific to our own setup and was not built with customizability in mind. That being said, we will share the code on request.

6. Conclusion

With this project we have implemented a simple, and easy-to-use sign language corpus for use as a lookup tool in learning situations in the classroom and beyond.

Since *STS-korpus* is intended as a tool mainly for novices and teachers of sign language, our focus has been simplicity of both purpose and design. We have therefore attempted to make *STS-korpus* easy to understand and use. Primarily this means the following:

- Removing the initial hurdle of having to download and install corpus data and ELAN or other similar software.
- Simplifying search and user-friendliness by minimizing clutter. Here we have imitated the appearance of the most commonly used web search engines, and instead of a complex interface with multiple fields and values, we display a single search box. Optional search prefixes (section 3.1.) can be used to perform advanced searches.
- Clearly highlighting the search results in their context. Search results are either shown as a concordance, or as a video with annotations, with search matches clearly highlighted in both cases.

More advanced features include:

- Access control (section 4.1.) by which registered users can be given access to additional corpora or annotation tiers.
- Dynamic content (section 4.3.) which teachers can use to upload their own annotated material for use in the classroom.

7. Acknowledgements

The research reported here was supported in part by the Swedish Sign Language Dictionary at the Department of Linguistics, Stockholm University. None of this would have been possible without the dedicated work of our colleagues Patrick Hansson and Thomas Björkstrand.

8. Bibliographical References

Crasborn, O. and Sloetjes, H. (2010). Using ELAN for Annotating Sign Language Corpora in a Team Setting. In P. Dreuw, et al., editors, *Proceedings of LREC 2010, Fourth Workshop on the Representation and Processing of Sign Languages*, pages 61–64, Valletta, Malta.

ELAN. (2020). [Computer Software]. (Version 5.9) Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/.

Granath, S. (2009). Who Benefits from Learning How to Use Corpora? In K. Aijmer, editor, *Corpora and Language Teaching*, volume 33 of *Studies in Corpus Linguistics*, pages 47–65. John Benjamin Publishing.

Hanke, T. (2002). iLex – A tool for Sign Language Lexicography and Corpus Analysis. In M. González Rodríguez et al., editors, *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation*, pages 923–926.

Leeson, L., Fenlon, J., Mesch, J., Grehan, C., and Sheridan, S. (2019). The Uses of Corpora in L1 and L2/Ln Sign Language Pedagogy. In R.S. Russell, editor, *The Routledge Handbook of Sign Language Pedagogy*, pages 339–352. Routledge, New York.

Mesch, J. and Wallin, L. (2008). Use of Sign Language Materials in Teaching. In *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*, pages 134–137.

Mesch, J. and Wallin, L. (2015). Gloss Annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1):102–120.

Wallin, L. and Mesch, J. (to be published 2020). *Annoteringskonventioner för teckenspråkstexter. Version 8.* [Annotation Conventions for Sign Language Discourse. Version 8]. Sign Language Section, Department of Linguistics, Stockholm University.

9. Language Resource References

Bergman, B. and Mesch, J. (2004). *Dataset. ECHO Dataset for Swedish Sign Language*. Department of Linguistics, Stockholm University.

Mesch, J., Wallin, L., Nilsson, A.-L., and Bergman, B. (2012). *Dataset. Swedish Sign Language Corpus Project 2009–2011 (Version 1).* Sign Language Section, Department of Linguistics, Stockholm University.

Mesch, J. (2016). *Dataset. Tactile Swedish Sign Language Corpus*. Sign Language Section, Department of Linguistics, Stockholm University.

Schönström, K. and Mesch, J. (2017). *Dataset. Corpus of Swedish Sign Language as Second Language Project 2013–2014 (Version 1).* Sign Language Section, Department of Linguistics, Stockholm University.

STS-korpus. (2020). [Swedish Sign Language Corpus Web Tool]. Sign Language Section, Department of Linguistics, Stockholm University. Retrieved February 5, 2020, from https://teckenspraksкorpus.su.se/.

Svenskt teckenspråkslexikon. (2008). [Swedish Sign Language Dictionary]. Sign Language Section, Department of Linguistics, Stockholm University. Retrieved February 5, 2020, from https://teckenspraksлexikon.su.se/.